

First-Person Hand Action Benchmark with RGB-D Videos and 3D Hand Pose Annotations

Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, Tae-Kyun Kim
Imperial College London, UK

{ggarciah, s.yuan14, s.baek15, tk.kim}@imperial.ac.uk

Abstract

In this work we study the use of 3D hand poses to recognize first-person hand actions interacting with 3D objects. Towards this goal, we collected RGB-D video sequences of more than 100K frames of 45 daily hand action categories, involving 25 different objects in several hand grasp configurations¹. To obtain high quality hand pose annotations from real sequences, we used our own mo-cap system that automatically infers the location of each of the 21 joints of the hand via 6 magnetic sensors on the finger tips and the inverse-kinematics of a hand model. To the best of our knowledge, this is the first benchmark for RGB-D hand action sequences with 3D hand poses. Additionally, we recorded the 6D (i.e. 3D rotations and locations) object poses and provide 3D object models for a subset of hand-object interaction sequences. We present extensive experimental evaluations of RGB-D and pose-based action recognition by 18 baselines/state-of-the-art. The impact of using appearance features, poses and their combinations are measured, and the different training/testing protocols including cross-persons are evaluated. Finally, we assess how ready the current hand pose estimation is when hands are severely occluded by objects in egocentric views and its influence on action recognition. From the results, we see clear benefits of using hand pose as a cue for action recognition compared to other data modalities. Our dataset and experiments can be of interest to communities of 6D object pose, robotics, and 3D hand pose estimation as well as action recognition.

1. Introduction

We, as humans, interact with the world using our hands to manipulate objects, machines, tools and socialize with other humans. In this work, we are interested in understanding how we use our hands in daily life actions using fine-grained hand pose features, a problem of interest for

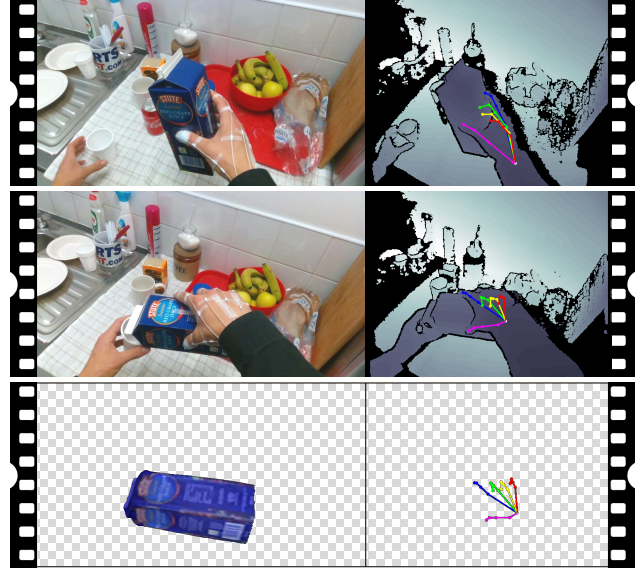


Figure 1: **Top two rows:** We show two frames belonging to the action class ‘pour juice’. In this work we propose a novel first-person action recognition dataset with RGB-D sequences and 3D hand pose annotations. We use magnetic sensors and inverse kinematics of a hand to capture the pose. On the right we see the captured depth image and hand pose. **Bottom row:** We captured 6D object pose for a subset of hand-object actions in our dataset to enable further research by the object pose community.

multiple applications requiring high precision such as hand rehabilitation [1], virtual reality [20] and robot imitation learning [2].

Previous work in first-person action recognition [6, 19, 25, 47] found that daily actions are well explained by looking at hands, a similar observation found in third-person view [66]. In these approaches, hand information is extracted from hand silhouettes [25, 47] or discrete grasp classification [6, 19, 42] using low-level image features. In [42] static actions and hand poses on synthetic data are provided,

¹Dataset visualization: <https://youtu.be/U5gleNWjz44>

whereas dynamic actions and hand poses on real sequences are provided in ours. In full-body human action recognition it is known that using higher level features such as body pose can benefit action recognition [43, 63, 68, 72]. Compared to full-body, hand actions present unique differences that make not obvious the use of pose as a cue: style and speed variations across subjects are more pronounced due to a higher degree of mobility of fingers and the motion can be very subtle. A setback for using hand pose for action recognition is the absence of reliable pose estimators off-the-shelf in contrast to full body [46, 62], mainly due to the absence of hand pose annotations on real (c.f. synthetic) data sequences notably when objects are involved [40, 41, 42].

In this work we introduce a new dataset of first-person hand action sequences with more than 100,000 RGB-D frames annotated with 3D hand poses, using six magnetic sensors attached to the finger tips and the use of inverse kinematics. We captured 1175 action samples including 45 categories manipulating 25 different objects in 3 scenarios. We designed our hand actions to cover multiple hand and grasp configurations and temporal dynamics. Furthermore, to encourage further research, we also provide 6-dimensional object pose ground truth (and their 3D mesh models) for 4 objects spanning 10 different actions. We evaluate several baselines and state-of-the-art RGB-D and pose-based action recognition in our dataset and test the current state-of-the-art in hand pose estimation and its influence on action recognition. To the best of our knowledge, this is the first work that studies the problem of first-person action recognition with the use of hand pose features and the first benchmark of its kind. In summary, the contribution of this paper is three-fold:

Dataset: we propose the first fully annotated dataset to help the study of egocentric hand actions and pose. This is the first dataset to combine both fields in the context of hands in real sequences and quality hand pose labels.

Action recognition: we evaluate 18 baselines and state-of-the-art approaches in RGB-D and pose-based action recognition using our proposed dataset. Our selected methods cover most of the research trends in both methodology and use of different data modalities.

Hand pose: We evaluate the state-of-the-art hand pose estimator in our real dataset i.e. the occluded setting of hand-object manipulations and evaluate its performance for action recognition.

2. Related work

Egocentric vision and manipulations datasets: The leading role of hands while manipulating objects has attracted the interest from both computer vision and robotics communities. From an action recognition perspective and only using RGB cues, recent work [3, 9, 10, 25, 36, 47] has delved into recognizing daily actions and discovered that

both object and hands are important cues to the recognition problem. Another related line of work is the study of human grasp from a robotics perspective [4, 5], as a cue for action recognition [6, 12, 19, 66], force estimation [12, 42] and as a recognition problem itself [16, 42]. In these previous works, hands are modeled using low-level features or intermediate representations following empirical grasp taxonomies [4] and thus limited compared to the 3D hand pose sequences used in this work. From a hand pose perspective, [40] proposed a small synthetic dataset of static poses and thus limited to train recent data-hungry algorithms. Given that we also provide 6-dimensional object poses and 3D mesh models for a subset of objects, our dataset can be of interest to both object pose and joint hand-object tracking emerging communities [48, 53]. To the best of our knowledge there is no such available egocentric hand-object dataset with both hand and object pose in 3D. We compare our dataset with other egocentric datasets in Section 3.5.

RGB-D and pose-based action recognition: Using depth sensors for human action recognition differs from traditional color action recognition in the fact that most successful color approaches [11, 58] cannot be directly applied to the depth stream due to its nature: noisy, textureless and discontinuous pixel regions led to the necessity of depth-tailored methods. These methods usually focus on how to extract discriminative features from the depth images using local geometric descriptors [32, 35, 65] sensitive to viewpoint changes and view-invariant approaches [38, 39]. However, the recent trend is to take advantage of the depth channel to obtain robust body pose estimates [46] and use them directly as a feature to recognize actions. This trend has led to what is known as pose (or skeleton) action recognition, a problem which is usually tackled as a time series problem. Popular approaches include the use of temporal state-space models [13, 61, 63, 64, 72], key-poses [57, 71], hand-crafted pose features [55, 56] and temporal recurrent models [8, 54, 73]. Having multiple data streams has led to the study of combining different sources of information such as depth and pose [32, 44, 60], color and pose [74] and all of them [15]. Most previous work in RGB-D action recognition focuses on actions performed by the whole human body with some exceptions that are mainly application-oriented such as hand gestures for human-computer interaction [7, 14, 24, 32] and sign language [59]. More related to us, [27] mounted a depth sensor to recognize egocentric activities and modeling hands using low-level skin features. Similar to our interests but in third-person view, [22, 67] used a hand tracker to obtain noisy estimates of hand pose in kitchen manipulation actions, while [7] recognized basic hand gestures for human-computer interaction without objects involved. In these works, actions performed and pose labels are very limited due to the low quality of the hand tracker, while in this work we provide accurate hand

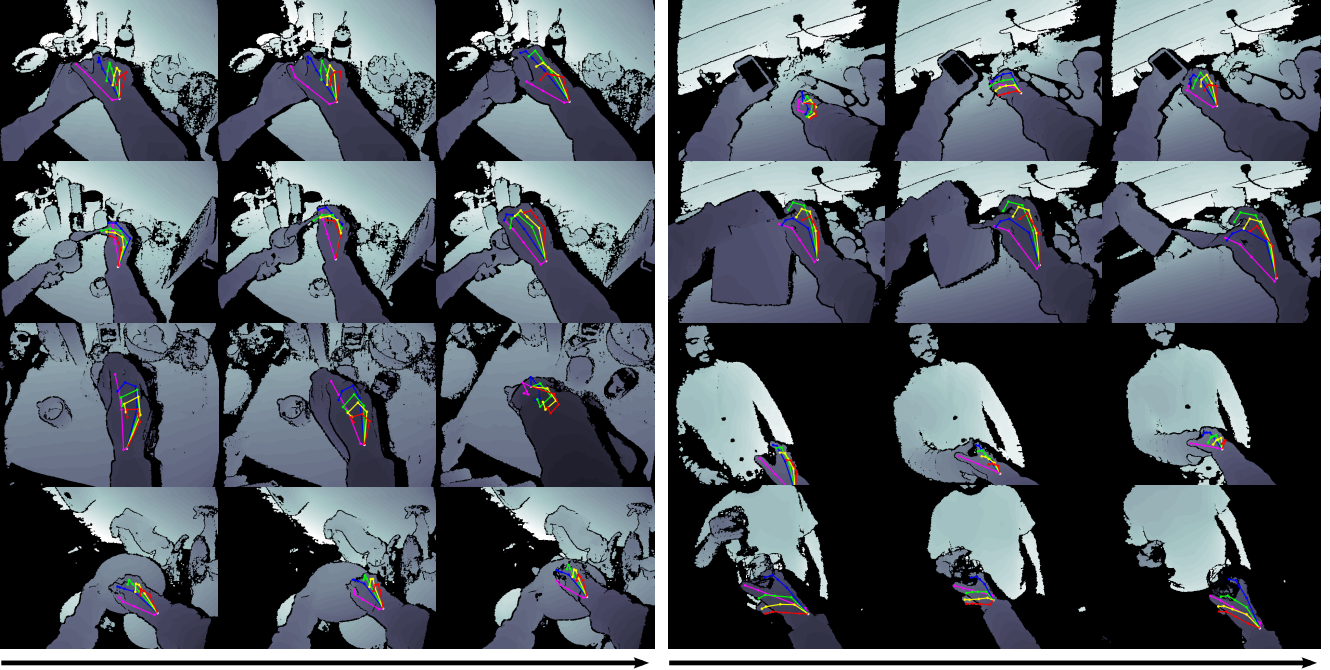


Figure 2: **Hand actions:** We have captured daily hand actions using a RGB-D sensor and used mo-cap to annotate hand pose. **Left** - from top to bottom: ‘open peanut butter’, ‘put sugar’, ‘pour milk’ and ‘wash with sponge’ (all in kitchen). **Right** - from top to bottom: ‘charge cell phone’ and ‘tear paper’ (office); ‘handshake’ and ‘toast with wine glass’ (social).

pose labels to study more realistic hand actions. We go in depth and evaluate various baselines and state-of-the-art approaches in Sections 4 and 5.

Hand pose estimation: Mainly due to the availability of commodity RGB-D sensors, hand pose estimation has recently made significant progresses in third-person view [18, 21, 23, 29, 31, 33, 37, 45, 51, 69] and more modest advances in first-person setting [30, 40, 41]. In [34], 3D tracking of a hand interacting with an object in their-person view was investigated. One main limitation of the field is the difficulty of obtaining accurate pose annotations leading researchers to resort to either synthetic [40, 45], manually or semi-automatically annotated [30, 49, 50, 52] datasets, resulting in not realistic images or low in number and often inconsistent labels respectively. Recently, and with the help of magnetic sensors for annotation, [70] has proposed a million-scale benchmark including egocentric poses (with no objects involved) and shown that a ConvNet baseline can obtain state-of-the-art results even in cross-dataset experiments when enough training data is available.

3. Daily hand-object actions dataset

In this section, we describe our Daily Hand-Object Actions dataset and we show a variety of relevant statistics.

3.1. Dataset overview

The dataset contains 1,175 action videos belonging to 45 different action categories in 3 different scenarios and performed by 6 actors. A total of 105,459 RGB-D frames are annotated with accurate hand pose and action category. Action sequences present high both inter-subject and intra-subject variability of style, speed, scale and viewpoint. Object 6-dimensional (location and angle in 3D) pose and mesh model are also provided for 4 objects involving 10 different action categories. Our plan is to keep growing the dataset with more models and objects. In Fig. 2 we show some example frames for different action categories and hand-pose annotation visualization.

3.2. Hand-object actions

We captured 45 different daily hand action categories involving 25 different objects (Fig. 4 (a)). We designed our action categories to span a high number of different grasp configurations and be diverse in both hand pose and action space (see Fig. 3). Each object has a minimum of one associated action (e.g. pen-‘write’) and a maximum of four (e.g. sponge-‘wash’, ‘scratch’, ‘squeeze’ and ‘flip’). These 45 hand actions were recorded and grouped in three different scenarios: kitchen (25), office (12) and social (8). Kitchen scenario (Fig. 2 left) comprises actions such as ‘stir’, ‘sprinkle’, ‘prick’ and ‘pour’, while some of the of-

fice actions (Fig. 2 top-right) include ‘write’, ‘type’ and ‘tear paper’. The social scenario (Fig. 2 bottom-right) contains interactions with other humans such as ‘handshake’, ‘high five’ and ‘toast with a glass of wine’. We also provide 6-dimensional object pose and mesh models for the following objects: ‘milk bottle’, ‘salt’, ‘juice carton’ and ‘liquid soap’. These objects are involved in 10 different hand-object action categories in the kitchen scenario.

In this work we have considered each hand-object manipulation as a different action category (similar to previous datasets [10]), although other definitions are plausible. For example, one could label ‘open juice carton’ and ‘open peanut butter’ as same action ‘open’ and/or make grammar combinations [67] of nouns (objects) and verbs (actions).

3.3. Sensors and data acquisition

Visual data: To capture visual data, we mounted on the shoulder of the subject the most recent version of the Intel RealSense SR300 RGB-D camera and captured RGB and Depth streams in the highest possible resolutions (1920×1080 and 640×480 for the color and depth stream respectively). We used the same frame rate of 30 fps in both streams.

Pose annotation: To obtain quality annotations of hand and object pose, we follow the approach of [70]. Hand pose is captured using six magnetic sensors [28] attached to the user’s hand (five fingertips and one wrist). Each sensor provides position and orientation with 6 degrees of freedom and the full hand pose is inferred using inverse kinematics over a defined 21-joint hand model. Each sensor is 2 mm wide and when attached to the human hand does not influence the depth image. The color image is affected as the sensors and the tape to attach them can be seen, however the hand is fully visible and actions distinguishable by using the color image (see Fig. 1). More details about the capture system can be found in the supplementary material and in [70]. For the object pose, we attach one another sensor to the closest point we can reach to the center of mass.

Recording process: We asked 6 people, 4 males and 2 females, all right-handed to perform the actions while having the mo-cap sensor and the camera attached. Instructions on how to perform the action in a safe manner (for the sensitive attached sensors) were given, however no instructions about style or speed were provided in order to capture realistic data. We found difficulties while acquiring the data due to the magnetic nature of the sensor as any metallic object would interfere and make the hand pose impossible to recover. This limited the objects to use and, in some cases, we resorted to their plastic versions (*e.g.* fork and spoon). We annotated the action labels manually.

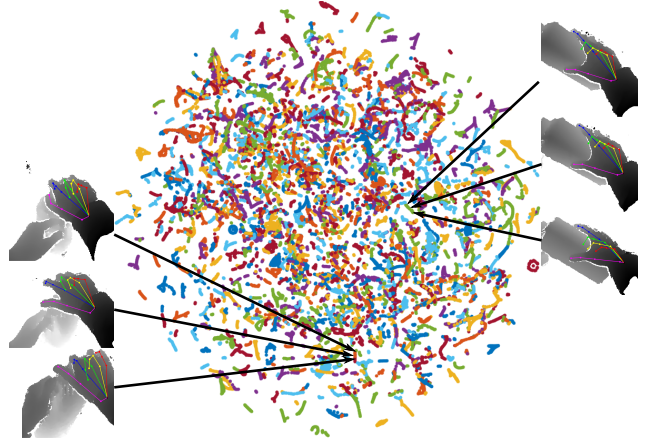


Figure 3: t-SNE [26] visualization of hand pose embedding over our dataset. Each colored dot represents a full hand pose and each trajectory an action sequence. Our dataset is rich in both hand pose configurations and actions space.

3.4. Dataset statistics

Fig. 4 (a) shows the distribution of different actions per involved object. Some objects such as spoon have multiple actions (‘stir’, ‘sprinkle’, ‘scoop’, ‘put sugar’) while some objects have one specific action (‘use calculator’). For completeness and although it is not an object, we include ‘hand’ as an object in actions ‘handshake’ and ‘high five’.

In Fig. 4 (b) we show the number of videos for each action class. On average there are 26.11 sequences per class action and 45.19 sequences per object category. The percentage of samples per scenario over all samples relates with the number of action categories per scenario.

Fig. 4 (c) shows the average number of frames per sequence for the 45 action classes. We observe that some action classes such as ‘put sugar’ and ‘open wallet’ involve short atomic movements (on average one second) while others such as ‘open letter’ require more time to be executed.

3.5. Comparison with other datasets

In Table 1 we summarize popular egocentric datasets that involve hands and objects in both dynamic and static fashion depending on their problem of interest. For conciseness, we have excluded from the table related datasets that do not partially or fully contain objects manipulations (*e.g.* [30, 36, 70]). Note that previous datasets in action recognition [3, 10] do not include hand pose labels. On the other hand, pose and grasp datasets [4, 5, 40, 42] do not contain dynamic actions and hand pose annotation is obtained by generating synthetic images. Our dataset ‘fills the gap’ of egocentric hand action using pose and compares favorably in terms of diversity, number of frames and use of real data.

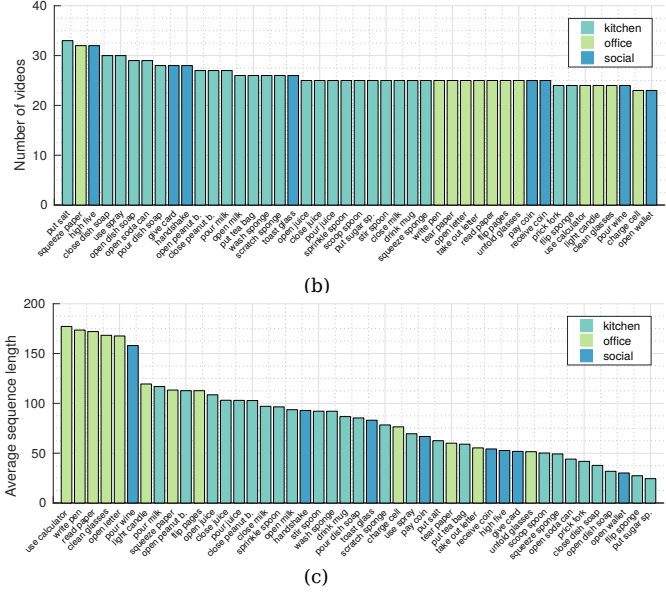
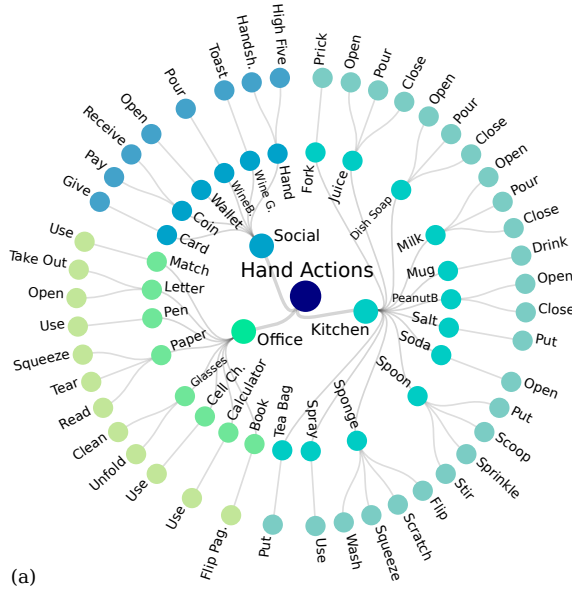


Figure 4: **(a)** Visualization of our hand actions involving objects dataset. Some objects are associated with multiple actions (e.g. spoon, sponge, liquid soap) while some others have only one linked action (e.g. calculator, pen, cell charger). **(b)** Number of action instances per hand action category. **(c)** Average number of frames in each video per hand action category. Our dataset contains both atomic and more temporally complex action classes.

Dataset	Cam.	Real?	Class.	Seq.	Frames	Labels
Yale [4]	RGB	✓	33	-	9100	Grasp
UTG [5]	RGB	✓	17	-	-	Grasp
GTEA [10]	RGB	✓	61	525	31,222	Act.
EgoHands [3]	RGB	✓	4	48	4,800	Act.
UCI-EGO [40]	RGB-D	✗	-	-	400	Pose
GUN-71 [42]	RGB-D	✓	71	-	12,000	Grasp
Ours	RGB-D	✓	45	1175	105,459	Act.+Pose

Table 1: First-person view datasets with hands and objects involved. Our proposed dataset is the first providing both pose and action labels.

4. Evaluated algorithms and baselines

4.1. Action recognition

In order to evaluate the current state-of-the-art in action recognition we chose a variety of approaches that, we believe, cover the most representative trends in the literature (Table 4). As the nature of our data is RGB-D and we have hand pose, we focus our attention to RGB-D and pose-based action recognition approaches, although we also evaluate two RGB action recognition methods [11, 15]. Note that as discussed above, most of previous work in RGB-D action recognition involve full body poses instead of hands and some of them might not be tailored for hand actions, we elaborate further on this in Section 5.1.

We start with one baseline to assess how the current state-of-the-art in color action recognition performs in our

dataset. For this, and given that most successful (color) action recognition approaches [25, 47] use ConvNets to learn descriptors from color and motion flow, we evaluate a recent two-stream architecture fine-tuned on our dataset [11].

About the depth modality, we first evaluate two local depth descriptor approaches HOG² [32] and HON4D [35] that exploit gradient and surface normal information as a feature for action recognition. As a global-scene depth descriptor, we evaluate the recent approach by [39] that learns view invariant features using ConvNets from several synthesized depth views of human body pose.

We follow our evaluation with pose-based action recognition methods. As our main baseline, we implemented a recurrent neural network with long-short term memory (LSTM) modules inspired in the architecture by [73]. We also evaluate several state-of-the-art pose action recognition approaches. We start with descriptor-based methods such as Moving Pose [71] that encodes atomic motion information and [55] who represents poses as points on a Lie group. For methods focusing on learning temporal dependencies, we evaluate HBRNN [8], Gram Matrix [72] and TF [13]. HBRNN consists of a bidirectional recurrent neural network with hierarchical layers designed to learn features from the body pose. Gram Matrix is currently the best performing method for body pose and uses Gram matrices to learn the dynamics of actions. TF proposes an approach that learns discriminative transitions between poses.

To conclude, we evaluate one hybrid approach that

jointly learns heterogeneous features (JOULE) [15]. JOULE uses a three-step iterative optimization algorithm to learn features jointly taking into account all the data channels (color, depth and hand pose). We use this particularity to study the contribution of different channels to action recognition in Section 5.1.

4.2. Hand pose estimation

To evaluate the state-of-the-art hand pose estimation we use the same ConvNet architecture as [70]. We choose this approach as it is easy to interpret and provided the best performance in the cross-benchmark evaluation by [70]. The details of the architecture are provided in the supplementary material. The chosen method is a discriminative approach operating on a frame-by-frame basis, which does not need any initialization and manual recovery when it fails in tracking [17, 33]. Most existing methods focus on hands alone, while some tracking-based methods deal with occlusions in hand-object interactions [34]. We report the accuracies of the ConvNet method trained with/without including objects (Section 5.2).

5. Evaluation results

5.1. Action recognition

In the following we present our experiments in action recognition. In this section we assume the hand pose is given, i.e. we use the hand pose annotations obtained using the magnetic sensors and inverse-kinematics. We evaluate the use of estimated hand poses without the aid of the sensors for action recognition in Section 5.2. As a measure of performance, we use the total accuracy of predicted actions or, in other words, correct number of predictions over total number of predictions. When possible we used publicly available codes with default parameters.

Hand pose pre-processing: Following the standard practice in body-pose action recognition [55, 71], we compensate for anthropomorphic differences by normalizing the hand poses to all have the same distance between pairs of joints. Furthermore, to be invariant to viewpoints, we define the center of coordinates to be the hand wrist. We found in our experiments that this normalization is important to obtain acceptable results.

5.1.1 A baseline: LSTM

We start our experimental evaluation with a simple yet powerful baseline: a recurrent neural network with long-short term memory module (LSTM). The architecture of our network is inspired by [73] with two differences: we do not ‘go deep’, and use a more conventional unidirectional network instead of bidirectional. Following [73], we set the number of neurons to 100 and a probability of dropout of

Protocol	1:3	1:1	3:1	cross-person
Acc. (%)	58.75	78.73	84.82	62.06

Table 2: Results for different training-testing protocols. 3:1 stands for ‘75% of the dataset is used for training and 25% for testing’. In cross-person protocol perform 6-fold leave-one-person-out cross-validation.

0.2. We use TensorFlow and Adam optimizer. We feed the normalized hand poses sequences into the LSTM.

Training and testing protocols: We experiment with two protocols. The first protocol consists of using different partitions of the data for training and the rest for testing and we tried three different training:testing ratios of 1 : 3, 1 : 1 and 3 : 1. The second protocol is a 6-fold ‘leave-one-person-out’ cross-validation, i.e. each fold consists of 5 subjects for training and one for testing. Results are presented in Table 2. We observe that following a cross-person protocol yields the worst results taking into account that in each fold we have similar training/testing proportions to the 3 : 1 setting. This can be explained by the difference in hand action styles between subjects. In the rest of the paper we perform our experiments using the 1:1 setting with 600 action sequences for training and 575 for testing. The result for this protocol is 78.73% using 1-layer LSTM. We also experimented adding with more layers, for example using 2-layer LSTM the accuracy improved to 80.14%. We did not observe any significant improvements by stacking more layers likely due to the given size of our dataset.

Results discussion: In Fig. 5 (a) we show the recognition accuracies per category on a subset actions. Some actions such as ‘sprinkle spoon’ ‘put tea bag’ and ‘pour juice’ are easily identifiable, while actions such as ‘open wallet’ and ‘use calculator’ are commonly confused, likely because hand poses are dissimilar and more subtle.

5.1.2 State-of-the-art

In Table 4 we show results for state-of-the-art approaches in different data modalities. We observe that the Two-stream [11] method performs well when combining both spatial and temporal cues. We observe that depth methods tend to perform slightly worse than the rest of the methods, suggesting that they are not able to fully capture neither the object cues nor the hand pose. Note that for Novel View [39] we extracted deep features from a network trained on several synthetic views of bodies, which might not generalize well to hand poses (fine-tuning in our dataset did not help). From all approaches, we observe that the ones using hand pose are the ones that achieve the best performance, with Gram Matrix [72] and Lie group [55] performing particularly well, a result in line with the ones reported in body

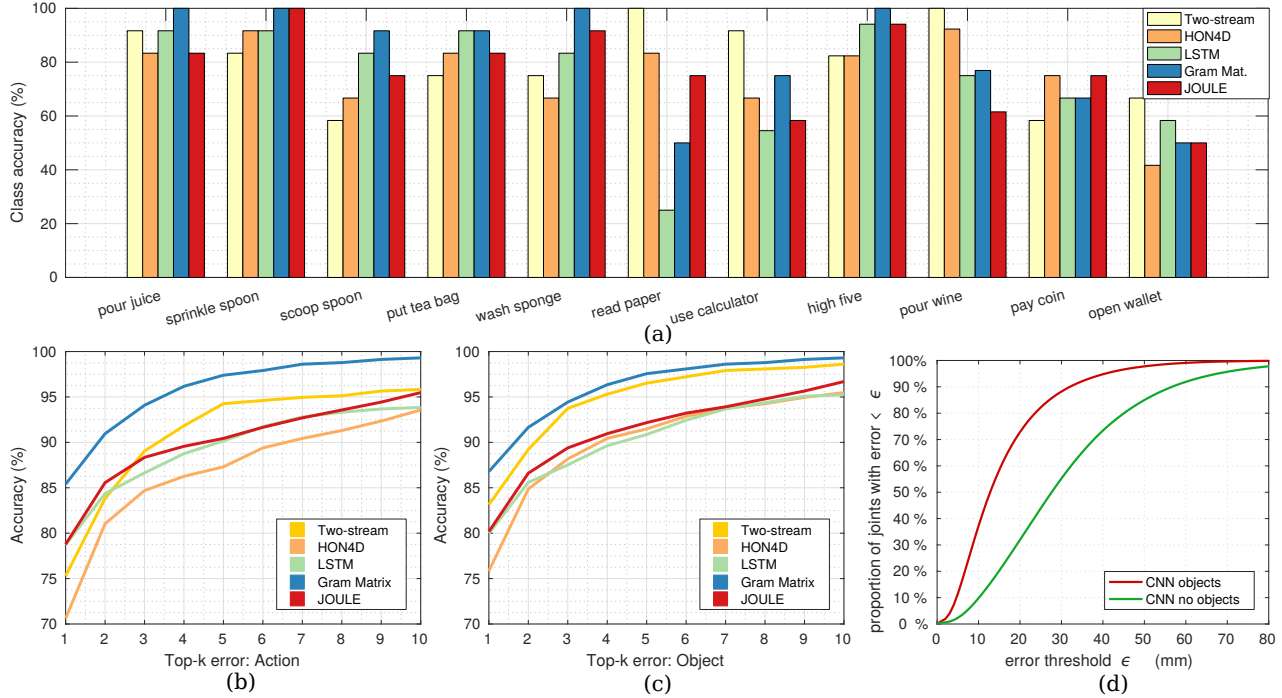


Figure 5: **(a)** We show class accuracies of some representative methods for different data modalities on a subset of classes. **(b)** Top- k action accuracy: true action label is in the top- k action prediction hypothesis. **(c)** Top- k object accuracy: manipulated object is in the top- k action prediction hypothesis. **(d)** Hand pose estimation error by training with objects/no objects.

pose action recognition.

In Fig. 5 we select some of the most representative methods and analyze their performance in more detail. We observe that the pose method Gram Matrix outperforms the rest in most of the measures, specially when we retrieve the top k action hypothesis (Fig. 5 (b)), showing the benefit of using hand pose for action recognition. Looking at Fig. 5 (a), we observe that Two-stream outperforms the rest of methods in some categories in which the object is big and the action does not involve much motion *e.g.* ‘use calculator’ and ‘read paper’. This good performance can be due to the pre-training of the spatial network on a big image recognition dataset. We further observe this in Fig. 5 (c) where we analyze the top k hypothesis given by the prediction and look whether the predicted action contains the object being manipulated, suggesting that the network correctly identifies the object but its temporal precision is worse than other methods.

Hand pose vs. depth vs. color: We performed one additional experiment using the JOULE approach by breaking down the contributions of each data modality. In Table 4 (bottom) we show that hand pose features are the most discriminative ones, although the performance can be increased by combining them with RGB and depth cues. This result suggests that hand poses provide complementary in-

Pose feature	Hand	Object	Hand+Object
Action acc. (%)	87.45	74.45	91.97

Table 3: We evaluate the use of 6D object pose for action recognition on a subset of our dataset. We observe the benefit of combining them with the hand pose.

formation to RGB and depth cues as previously observed in body pose action recognition.

Object pose: We experimented using the object pose as an additional feature for action recognition using the subset of actions that have annotated object poses: a total of 261 sequences for 10 different classes and 4 objects. We trained our LSTM baseline on half of the sequences and using three different inputs: hand pose, object pose and both combined. In Table 3 we show the results and observe that combining both pose features can help the action recognition.

5.2. Hand pose estimation

Input pre-processing: Most approaches in hand pose estimation require to have as an input the depth channel containing only the hand. To crop hands, we used the bounding boxes automatically annotated using the magnetic sensors. Detecting hands in depth images is still an open problem

Method	Year	Color	Depth	Pose	Acc (%)
Two stream-color [11]	2016	✓	✗	✗	61.56
Two stream-flow [11]	2016	✓	✗	✗	69.91
Two stream-all [11]	2016	✓	✗	✗	75.30
HOG ² -depth [32]	2013	✗	✓	✗	59.83
HOG ² -depth+pose [32]	2013	✗	✓	✓	66.78
HON4D [35]	2013	✗	✓	✗	70.61
Novel View [39]	2016	✗	✓	✗	69.21
1-layer LSTM	2016	✗	✗	✓	78.73
2-layer LSTM	2016	✗	✗	✓	80.14
Moving Pose [71]	2013	✗	✗	✓	56.34
Lie Group [55]	2014	✗	✗	✓	82.69
HBRNN [8]	2015	✗	✗	✓	77.40
Gram Matrix [72]	2016	✗	✗	✓	85.39
TF [13]	2016	✗	✗	✓	80.69
JOULE-color [15]	2015	✓	✗	✗	66.78
JOULE-depth [15]	2015	✗	✓	✗	60.17
JOULE-pose [15]	2015	✗	✗	✓	74.60
JOULE-all [15]	2015	✓	✓	✓	78.78

Table 4: Hand action recognition performance by different evaluated approaches on our proposed dataset.

that we do not investigate further. As future work we plan to use one hand detector, for instance in [42], to estimate the input to the hand pose estimator. The quality of the detection is likely to affect the hand pose estimation.

Training with objects vs. no objects: One question raised while designing our experiments was whether we actually needed to annotate the hand pose in a close to ground truth accuracy, to experiment with hand actions. We try to answer this question by estimating the hand poses of our hand action dataset in two ways: using the nearly 300k *object-free* egocentric samples from [70] and using the images in the training set of our hand action dataset. As observed in Fig. 5 (d), the results suggest that having hand-object images in the training set is crucial to train state-of-the-art hand pose estimators likely due to the fact that occlusions and object shapes need to be *seen* by the estimator beforehand. This shows the need of having annotated hand poses in manipulation hand poses and thus why our dataset can be of interest for the hand pose community. In future, generalization/scalability issues of the annotated hand pose with objects will be further investigated.

Hand pose estimation and action recognition: Now we try to answer the following key question: ‘how good is the current hand pose estimation for recognizing hand actions?’. In Table 5 we show results of hand action recognition by swapping the hand pose labels by the estimated ones in the testing set. We observe that reducing the hand pose error in a factor of two yields a more than twofold improvement in action recognition. The difference in hand action recognition between using the hand pose labels in testing and using the estimated ones is 6.67%. While

Training set	Pose error (mm)	Action Acc. (%)
Ours	14.34	72.06
[70]	31.03	29.63
Magnetic sensors and IK	-	78.73

Table 5: Hand pose estimation error and its impact on hand-object action recognition. Improving the hand pose estimation yields to best action recognition performance.

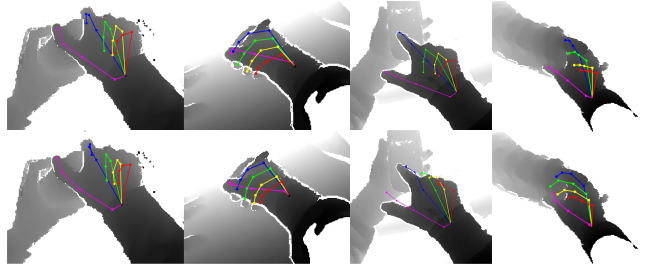


Figure 6: **Top row:** pose labels obtained using the magnetic sensors. **Bottom row:** hand pose estimates. Some estimates are noisy but good enough for action recognition.

this difference is significant, using estimated poses and our LSTM baseline we can achieve a comparable or higher performance to other RGB-D approaches without using hand pose. We also tested the two best performant state-of-the-art methods from previous section, Lie group [55] and Gram Matrix [72]. For Lie group we obtained an accuracy of 69.22%, while for Gram Matrix a poor result of 32.22% likely due to strong assumptions in the noise distribution of their model. On the other hand, our LSTM baseline shows more robust behavior in the presence of noisy hand pose estimates. In Fig. 6 we show some qualitative results in hand pose estimation in our proposed dataset.

6. Concluding remarks

We have proposed a novel benchmark and presented experimental evaluations for RGB-D and pose-based, hand action recognition, in first-person settings. The benchmark provides both temporal action labels and full 3D hand pose labels, and additionally 6D object pose labels on a part of the dataset. Both RGB-D action recognition and 3D hand pose estimation are relatively new fields, and this is a first attempt to relate both of them as happened for full human body. We have evaluated several baselines in our dataset and concluded that hand pose features are a rich source of information for recognizing manipulation actions. We believe that our dataset and experiments can encourage future work in multiple fields including action recognition, hand pose estimation, object pose estimation, grasp analysis and emerging ones such as joint hand-object pose estimation.

References

- [1] S. Allin and D. Ramanan. Assessment of post-stroke functioning using machine vision. In *MVA*, 2007.
- [2] B. D. Argall, S. Chernova, M. Veloso, and B. Browning. A survey of robot learning from demonstration. In *RSS*, 2009.
- [3] S. Bambach, S. Lee, D. J. Crandall, and C. Yu. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In *ICCV*, 2015.
- [4] I. M. Bullock, T. Feix, and A. M. Dollar. The yale human grasping dataset: Grasp, object, and task data in household and machine shop environments. In *International Journal of Robotics Research*, 2015.
- [5] M. Cai, K. M. Kitani, and Y. Sato. A scalable approach for understanding the visual structures of hand grasps. In *ICRA*, 2015.
- [6] M. Cai, K. M. Kitani, and Y. Sato. Understanding hand-object manipulation with grasp types and object attributes. In *RSS*, 2016.
- [7] Q. De Smedt, H. Wannous, and J.-P. Vandeboorde. Skeleton-based dynamic hand gesture recognition. In *CVPRW*, 2016.
- [8] Y. Du, W. Wang, and L. Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *CVPR*, 2015.
- [9] A. Fathi, A. Farhadi, and J. M. Rehg. Understanding egocentric activities. In *ICCV*, 2011.
- [10] A. Fathi, X. Ren, and J. M. Rehg. Learning to recognize objects in egocentric activities. In *CVPR*, 2011.
- [11] C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutional two-stream network fusion for video action recognition. In *CVPR*, 2016.
- [12] C. Fermüller, F. Wang, Y. Yang, K. Zampogiannis, Y. Zhang, F. Barranco, and M. Pfeiffer. Prediction of manipulation actions. In *IJCV*, 2017.
- [13] G. Garcia-Hernando and T.-K. Kim. Transition forests: Learning discriminative temporal transitions for action recognition and detection. In *CVPR*, 2017.
- [14] P. M. X. Y. S. Gupta and K. K. S. T. J. Kautz. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural networks. In *CVPR*, 2016.
- [15] J.-F. Hu, W.-S. Zheng, J. Lai, and J. Zhang. Jointly learning heterogeneous features for RGB-D activity recognition. In *CVPR*, 2015.
- [16] D.-A. Huang, M. Ma, W.-C. Ma, and K. M. Kitani. How do we use our hands? discovering a diverse set of common grasps. In *CVPR*, 2015.
- [17] Intel. Perceptual computing sdk. 2013.
- [18] C. Ionescu, J. Carreira, and C. Sminchisescu. Iterated second-order label sensitive pooling for 3d human pose estimation. In *CVPR*, 2014.
- [19] T. Ishihara, K. M. Kitani, W.-C. Ma, H. Takagi, and C. Asakawa. Recognizing hand-object interactions in wearable camera videos. In *ICIP*, 2015.
- [20] Y. Jang, S.-T. Noh, H. J. Chang, T.-K. Kim, and W. Woo. 3d finger cape: Clicking action and position estimation under self-occlusions in egocentric viewpoint. In *IEEE TVCG*, 2015.
- [21] C. Keskin, F. Kırac, Y. E. Kara, and L. Akarun. Hand pose estimation and hand shape classification using multi-layered randomized decision forests. In *ECCV*, 2012.
- [22] J. Lei, X. Ren, and D. Fox. Fine-grained kitchen activity recognition using rgb-d. In *Proc. ACM Ubiquitous Computing*, 2012.
- [23] H. Liang, J. Yuan, and D. Thalmann. Parsing the hand in depth images. In *TMM*, 2014.
- [24] L. Liu and L. Shao. Learning discriminative representations from rgb-d video data. In *Proc. IJCAI*, 2013.
- [25] M. Ma, H. Fan, and K. M. Kitani. Going deeper into first-person activity recognition. In *CVPR*, 2016.
- [26] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. In *Journal of Machine Learning Research*, 2008.
- [27] M. Moghimi, P. Azagra, L. Montesano, A. C. Murillo, and S. Belongie. Experiments on an rgb-d wearable vision system for egocentric activity recognition. In *CVPRW*, 2014.
- [28] NDITrakSTAR. <http://www.ascension-tech.com/products/trakstar-2-drivebay-2/>.
- [29] N. Neverova, C. Wolf, G. W. Taylor, and F. Nebout. Hand segmentation with structured convolutional learning. In *ACCV*, 2014.
- [30] M. Oberweger, G. Riegler, P. Wohlhart, and V. Lepetit. Efficiently creating 3d training data for fine hand pose estimation. In *CVPR*, 2016.
- [31] M. Oberweger, P. Wohlhart, and V. Lepetit. Training a feedback loop for hand pose estimation. In *ICCV*, 2015.
- [32] E. Ohn-Bar and M. M. Trivedi. Hand gesture recognition in real time for automotive interfaces: A multimodal vision-based approach and evaluations. In *ITS*, 2014.
- [33] I. Oikonomidis, N. Kyriazis, and A. A. Argyros. Efficient model-based 3d tracking of hand articulations using kinect. In *BMVC*, 2011.
- [34] I. Oikonomidis, N. Kyriazis, and A. A. Argyros. Full dof tracking of a hand interacting with an object by modeling occlusions and physical constraints. In *ICCV*, 2011.
- [35] O. Oreifej and Z. Liu. HON4D: histogram of oriented 4D normals for activity recognition from depth sequences. In *CVPR*, 2013.
- [36] H. Pirsivash and D. Ramanan. Detecting activities of daily living in first-person camera views. In *CVPR*, 2012.
- [37] C. Qian, X. Sun, Y. Wei, X. Tang, and J. Sun. Realtime and robust hand tracking from depth. In *CVPR*, 2014.
- [38] H. Rahmani, A. Mahmood, D. Q. Huynh, and A. Mian. Histogram of oriented principal components for cross-view action recognition. *TPAMI*, 2016.
- [39] H. Rahmani and A. Mian. 3d action recognition from novel viewpoints. In *CVPR*, 2016.
- [40] G. Rogez, M. Khademi, J. Supančič III, J. M. M. Montiel, and D. Ramanan. 3d hand pose detection in egocentric rgb-d images. In *Proc ECCV*, 2014.
- [41] G. Rogez, J. S. Supancic, and D. Ramanan. First-person pose recognition using egocentric workspaces. In *CVPR*, 2015.
- [42] G. Rogez, J. S. Supancic, and D. Ramanan. Understanding everyday hands in action from rgb-d images. In *ICCV*, 2015.
- [43] A. Shahrudiy, J. Liu, T.-T. Ng, and G. Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *CVPR*, 2016.
- [44] A. Shahrudiy, T.-T. Ng, Q. Yang, and G. Wang. Multimodal multitask learning for action recognition in depth videos. In *IEEE TPAMI*, 2016.
- [45] T. Sharp, C. Keskin, D. Robertson, J. Taylor, J. Shotton, D. K. C. R. I. Leichter, A. V. Y. Wei, D. F. P. K. E. Krupka,

- A. Fitzgibbon, and S. Izadi. Accurate, robust, and flexible real-time hand tracking. In *CHI*, 2014.
- [46] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore. Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, 2013.
- [47] S. Singh, C. Arora, and C. Jawahar. First person action recognition using deep learned descriptors. In *CVPR*, 2016.
- [48] S. Sridhar, F. Mueller, M. Zollhöfer, D. Casas, A. Oulasvirta, and C. Theobalt. Real-time joint tracking of a hand manipulating an object from rgb-d input. In *ECCV*, 2016.
- [49] X. Sun, Y. Wei, S. Liang, X. Tang, and J. Sun. Cascaded hand pose regression. In *CVPR*, 2015.
- [50] D. Tang, H. J. Chang, A. Tejani, and T.-K. Kim. Latent regression forest: Structured estimation of 3d articulated hand posture. In *CVPR*, 2014.
- [51] D. Tang, T.-H. Yu, and T.-K. Kim. Real-time articulated hand pose estimation using semi-supervised transductive regression forests. In *ICCV*, 2013.
- [52] J. Tompson, M. Stein, Y. Lecun, and K. Perlin. Real-time continuous pose recovery of human hands using convolutional networks. In *TOG*, 2014.
- [53] D. Tzionas, L. Ballan, A. Srikantha, P. Aponte, M. Pollefeys, and J. Gall. Capturing hands in action using discriminative salient points and physics simulation. In *IJCV*, 2016.
- [54] V. Veeriah, N. Zhuang, and G.-J. Qi. Differential recurrent neural networks for action recognition. In *ICCV*, 2015.
- [55] R. Vemulapalli, F. Arrate, and R. Chellappa. Human action recognition by representing 3d skeletons as points in a lie group. In *CVPR*, 2014.
- [56] R. Vemulapalli and R. Chellappa. Rolling rotations for recognizing human actions from 3d skeletal data. In *CVPR*, 2016.
- [57] C. Wang, Y. Wang, and A. L. Yuille. Mining 3d key-pose-motifs for action recognition. In *CVPR*, 2016.
- [58] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *CVPR*, 2011.
- [59] J. Wang, Z. Liu, J. Chorowski, Z. Chen, and Y. Wu. Robust 3d action recognition with random occupancy patterns. In *ECCV*, 2012.
- [60] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *CVPR*, 2012.
- [61] P. Wang, C. Yuan, W. Hu, B. Li, and Y. Zhang. Graph based skeleton motion representation and similarity measurement for action recognition. In *ECCV*, 2016.
- [62] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *CVPR*, 2016.
- [63] D. Wu and L. Shao. Leveraging hierarchical parametric networks for skeletal joints based action segmentation and recognition. In *CVPR*, 2014.
- [64] L. Xia, C.-C. Chen, and J. Aggarwal. View invariant human action recognition using histograms of 3d joints. In *CVPRW*, 2012.
- [65] X. Yang and Y. Tian. Super normal vector for activity recognition using depth sequences. In *CVPR*, 2014.
- [66] Y. Yang, C. Fermuller, Y. Li, and Y. Aloimonos. Grasp type revisited: A modern perspective on a classical feature for vision. In *CVPR*, 2015.
- [67] Y. Yang, A. Guha, C. Fermuller, and Y. Aloimonos. A cognitive system for understanding human manipulation actions. In *Advances in Cognitive Systems*, 2014.
- [68] A. Yao, J. Gall, G. Fanelli, and L. J. Van Gool. Does human action recognition benefit from pose estimation?. In *BMVC*, 2011.
- [69] Q. Ye, S. Yuan, and T.-K. Kim. Spatial attention deep net with partial pso for hierarchical hybrid hand pose estimation. In *ECCV 2016*.
- [70] S. Yuan, Q. Ye, B. Stenger, S. Jain, and T.-K. Kim. Big-Hand2.2m benchmark: Hand pose dataset and state of the art analysis. In *CVPR*, 2017.
- [71] M. Zanfir, M. Leordeanu, and C. Sminchisescu. The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection. In *ICCV*, 2013.
- [72] X. Zhang, Y. Wang, M. Gou, M. Sznajder, and O. Camps. Efficient temporal sequence comparison and classification using gram matrix embeddings on a riemannian manifold. In *CVPR*, 2016.
- [73] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, and X. Xie. Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks. In *AAAI*, 2016.
- [74] Y. Zhu, W. Chen, and G. Guo. Fusing spatiotemporal features and joints for 3d action recognition. In *CVPRW*, 2013.